Pattern Recognition 44 (2011) 2375–2383

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Active learning with adaptive regularization

Zheng Wang [a,*], Shuicheng Yan [b], Changshui Zhang [a]

[a] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China

[b] Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117576, Singapore

ARTICLE INFO

Article history:
Received 15 September 2010
Received in revised form
18 January 2011
Accepted 7 March 2011
Available online 15 March 2011

Keywords:
Active learning
Adaptive regularization
SVM
TSVM

ABSTRACT

In classification problems, active learning is often adopted to alleviate the laborious human labeling efforts, by finding the most informative samples to query the labels. One of the most popular query strategy is selecting the most uncertain samples for the current classifier. The performance of such an active learning process heavily relies on the learned classifier before each query. Thus, stepwise classifier model/parameter selection is quite critical, which is, however, rarely studied in the literature. In this paper, we propose a novel active learning support vector machine algorithm with adaptive model selection. In this algorithm, before each new query, we trace the full solution path of the base classifier, and then perform efficient model selection using the unlabeled samples. This strategy significantly improves the active learning efficiency with comparatively inexpensive computational cost. Empirical results on both artificial and real world benchmark data sets show the encouraging gains brought by the proposed algorithm in terms of both classification accuracy and computational cost.

1. Introduction

Nowadays, with the fast development of Internet techniques and explosive increasing of data warehouses, the unlabeled data is becoming abundant or easily obtained in most cases. On the other hand, annotating the unlabeled samples is costly work and very time consuming. The label information, however, is very important for training a satisfactory learner in machine learning and other related problems. Therefore, more and more attention has been paid to finding a good classifier with minimum labeling efforts in recent years.

Active learning is one of the most popular techniques to save human labeling efforts. It has been widely adopted into the sophisticated supervised and semi-supervised tasks [12]. Active learning support vector machine (SVM) [14] is one of the most representative and practical approaches for pool based active learning in machine learning literature. The pioneering investigation is based on hard margin SVM. Furthermore, Campbell et al. [2] introduce active learning into the soft margin SVM, which is much more powerful and practical to deal with the nonseparable classes [15]. Both of these two algorithms prefer to choose the most uncertain sample for current classifier and query its label. This is a typical query criterion for the discriminant models, owing to its simplicity and efficiency [12]. In such active learning scenarios, the label query and classifier modeling are highly correlated. As the query heavily relies on the current classifier, an inappropriate model may lead to very poor active learning performance [9], which behaves as unsatisfactory learning accuracy and inefficient queries. Though active learning is well known for its benefit of saving labeling efforts, it is often less efficient than random query in the initial stages, with very few labeled data. This phenomena can be observed empirically in many popular active learning methods [1,14,17]. This might further prejudice the query efficiency of the whole active learning process, using such an unsatisfactory initialization. As a result, model selection is critical for active learning [13,1].

Choosing the best model is a very difficult problem in both machine learning and statistics fields [15]. It is often reduced and formulated as the parameter selection problem. In soft margin SVM, finding the best classifier can be formulated as a regularized optimization problem. A tunable parameter is used to control the regularization quantity. Given the loss function and the penalty, selection of a good value for the tunable parameter is the model selection problem [7]. In conventional supervised learning methods, the training data is often given beforehand and fixed. In this situation, once a satisfactory parameter is found, it will be fixed as a constant and used all through the following learning process. However, in active learning scenario, the number of the labeled data continually increases with the machine queries. During this process, the training data compose a dynamic set. Correspondingly, the

* Corresponding author.
E-mail address: wangzheng04@gmail.com (Z. Wang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.03.008

learning model should be changed with respect to the data set. In this work, we will show that using fixed regularization parameters is not a very good choice for active learning problems. When the number of labeled data increases, a dynamic parameter is desirable to guarantee a satisfactory learning result.

In this paper, we propose a very efficient active learning method for soft margin SVM with model selection. In supervised learning scenario, cross-validation is one of the most popular ways to find the proper parameter for the SVM model [15,7]. It is to split the available labeled data into a training and a validation sets. The training data is used to construct the SVM classifiers for different parameters, and the validation set is then used to select the most proper one. However, this setting cannot be easily adopted into the active learning framework. The most important reason is that the queried samples are not independently and identically distributed when sampled from the original data distribution. Using the queried data as the validation set may get severe overfitting and thus mislead the following query process. The problem becomes even worse when the original available labeled samples are scarce, which is commonly seen in the active learning. This is also one probable reason why conventional active learning is often less efficient than random query in the initial stages. To tackle this issue, in this work we use the unlabeled samples to compose a pseudo-validation set, and we prove that it works well both theoretically and empirically. To make the parameter selection process more efficient, we introduce the regularization path method [7] into the active learning process to efficiently compute the models based on different regularization values.

The rest of this paper is organized as follows. In Section 2, we introduce the model selection step into active learning framework based on SVM. In Section 3, we present a practical active learning algorithm with an adaptive model, which is called active learning SVM path. In Section 4, we discuss the relationship and difference between our proposed method and the conventional transductive SVM (TSVM) method. The experiments for empirical analysis are given in Section 5. Finally, we conclude in Section 6.

2. Active learning with dynamic SVM

Suppose initially there are l labeled points $\mathbf{X}_L=\{\mathbf{x}_1,\ldots,\mathbf{x}_l\}$, $\mathbf{x}_i \in \mathcal{X} \subseteq \mathcal{R}^m$, with labels $\mathbf{y}_L=\{y_1,\ldots,y_l\}$, where $y_i \in \mathcal{Y}=\{1,-1\}$ as we focus on binary problems here. There is also a pool of u unlabeled points $\mathbf{X}_U=\{\mathbf{x}_{l+1},\ldots,\mathbf{x}_{l+u}\}$, $\mathbf{D}_L=\{\mathbf{X}_L, \mathbf{y}_L\}$ are randomly generated according to some unknown probability $P(\mathbf{x},y)$. $\mathbf{D}_U=\mathbf{X}_U$ are randomly generated from the marginal probability $P(\mathbf{x})$.

2.1. Soft margin SVM formulation

The soft margin SVM searches for an optimal hyperplane $f(\mathbf{x})=\mathbf{w}^T\Phi(\mathbf{x})+b$ by solving the following optimization problem[1]:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i^s,$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\;\; \xi_i \geq 0,\;\; 1\leq i\leq l,$$

where C is a trade-off parameter and $\Phi(\mathbf{x}_i)$ is a function mapping the input data into a feature space where the data is better discriminated or represented. For linear case, $\Phi(\mathbf{x})=\mathbf{x}$.

In statistical learning theory, based on structural risk minimization, the soft margin SVM can also be formulated within the regularized unconstrained optimization framework as follows:

$$\hat{f}=\underset{f\in\mathcal{F}}{\operatorname{argmin}}\sum_{i=1}^{l}L(y_i,f(\mathbf{x}_i))+\lambda\Omega(f), \quad (1)$$

where $L(y_i,f(\mathbf{x}_i))=\max(0,1-y_if(\mathbf{x}_i))$ is the hinge loss function, $\Omega(f)=\|\mathbf{w}\|^2$ is the regularization term, which describes the model complexity, and λ is the regularization parameter, which controls the regularization quantity. Selection of a good value of λ is a so-called model selection problem.

The optimal solution for the soft margin SVM can be explicitly expressed as

$$\hat{f}(\mathbf{x})=\frac{1}{\lambda}\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x},\mathbf{x}_i)+b, \quad (2)$$

where $K(\mathbf{x},\mathbf{x}_i)=\Phi(\mathbf{x})\cdot\Phi(\mathbf{x}_i)$ is the kernel function. In this formulation, there are only a part of the data involved in the expression of the classifier, for which $0<\alpha_i\leq 1$.

2.2. Active learning SVM

In standard active learning SVM, the optimal classifier is among the hypotheses correctly classifying current labeled data. These consistent hypotheses compose the version space [14]. The active learner tries to find the new sample, which expectedly reduces the version space optimally, to query [14]. Though there is no consistent hypothesis to compose the version space in inseparable case for soft margin active learning SVM, the same query scenario is still found to be the most reasonable one [2]. This is best known as the most uncertainty principle, which is used in many active learning methods [12]. It can be expressed as

$$i=\underset{i\in U}{\operatorname{argmin}}|\hat{f}(\mathbf{x}_i)|. \quad (3)$$

The premise of query efficiency under this strategy is a well trained classifier $\hat{f}(\mathbf{x})$ in (2), which gives accurate predictions for unlabeled samples. It is constructed by the soft margin SVM, which is the solution of the regularized optimization, expressed by (1).

2.3. Active learning with dynamic regularization

The target of the learning problem is to find the optimal classifier, which is expectable to best approximate the Bayesian rule. Under this circumstance, active learning and model selection are two complementary and heavily correlated problems. However, these two parts have been studied separately as two independent problems, and little research has been done to consider them together. So far as we know, [13] is the only work to analyze such a problem. In [13], an ensemble active learning method is presented for linear regression, which averages all available models for active learning and is computationally expensive. Besides, the proposed algorithm in [13] cannot deal with the classification problem, which is thus still an open problem.

It is obvious that in soft margin SVM and other similar regularized optimization problems, the solution is decided by two factors. One is the current training set $\mathbf{D}_L=\{\mathbf{X}_L, \mathbf{y}_L\}$ and the other one is the regularization parameter λ. The optimum solution can be expressed as a function of these two factors, $\hat{f}(\cdot)=\hat{f}(\cdot|\mathbf{D}_L,\lambda)$. The target is to find the $\hat{f}$ to best approximate the Bayesian classifier, using all currently available knowledge. Intuitively, it is reasonable to use different amounts of regularization when training different numbers of labeled data. Proposition 1 preliminarily analyzes the behavior of the regularization parameter changing with respect to the increase of

[1] In this paper we focus on the 1-norm soft margin SVM with $s=1$.